

Randomization in Cancer Clinical Trials: Permutation Test and Development of a Computer Program

by Yasuo Ohashi*

When analyzing cancer clinical trial data where the treatment allocation is done using dynamic balancing methods such as the minimization method for balancing the distribution of important prognostic factors in each arm, conservativeness occurs if such a randomization scheme is ignored and a simple unstratified analysis is carried out. In this paper, the above conservativeness is demonstrated by computer simulation, and the development of a computer program that carries out permutation tests of the log-rank statistics for clinical trial data where the allocation is done by the minimization method or a stratified permuted block design is introduced. We are planning to use this program in practice to supplement a usual stratified analysis and model-based methods such as the Cox regression.

The most serious problem in cancer clinical trials in Japan is how to carry out the quality control or data management in trials that are initiated and conducted by researchers without support from pharmaceutical companies. In the final section of this paper, one international collaborative work for developing international guidelines on data management in clinical trials of bladder cancer is briefly introduced, and the differences between the system adopted in US/European statistical centers and the Japanese system is described.

Background

As seen in the guidelines by Simon and Wittes (1), high quality is being required in cancer clinical trial data for raising the reliability and comparability of trials, and this pressure from abroad is now influencing the design and management of cancer clinical trials in Japan. Until a few years ago, the randomization has been done almost exclusively by the envelop method in Japan, and a high rate of ineligibility and protocol violations have often deteriorated the reliability of the results. Recently, the central registration and randomization system by using a telephone call (or a facsimile) with a check of the eligibility criteria of each patient is rapidly spreading, and some medical researchers have a somewhat radical opinion that a trial by the envelop method is not scientifically evaluative.

If this centralized registration system is adopted and works well, it becomes possible to incorporate dynamic balancing methods for reducing a possible imbalance in the patients' distribution of important prognostic factors [see Kalish and Begg (2) for a review of proposed methods]. In many U.S./European cooperative phase III cancer clinical trials coordinated by well-organized statistical centers, the dynamic balancing is a usual practice rather than an exception, and the minimization

method (3) with some modifications is usually adopted (4). In Japan, Sakamoto et al. (5) reported that they are conducting a trial of gastrointestinal cancer with the minimization method.

Conservatism in Analysis

It is well known (6) that conservativeness occurs in analyzing clinical trial data if stratifying variables are ignored when the patient allocation is done by a stratified (block) design. Anderson (personal communication) and Forsythe and Stitt (7) pointed out that it also occurs for the minimization method. Kalish and Begg (8) studied the impact of the treatment allocation on nominal significance levels and concluded that "(nominal p -values) are not likely to be severely distorted if the analysis is stratified by important covariates used as allocation prompts." In many U.S./European centers that are using the minimization method, the stratified analysis is routine, and the conclusion by Kalish and Begg seems to support their strategy. The number of stratifying variables, however, is usually two or three and sometimes over four in U.S./European typical cancer clinical trials, and it also seems that the number of strata is sometimes over ten. It is easily expected that the efficiency (power) loss may not be negligible if unnecessary covariates are used in dynamic balancing; however, the research in this respect is not yet adequate. Moreover, in many U.S./European multiinstitutional cancer

*University Hospital Computer Center, University of Tokyo Hospital, Hongo, Bunkyo-Ku, Tokyo 113, Japan.

clinical trials, institutes are treated as different levels of covariate and incorporated in the minimization process. The above conservativeness will occur if they are ignored in analysis, and the efficiency loss will occur if the stratification is done as to institutions.

Table 1 is the result of our computer simulation that shows the conservativeness of the log-rank test statistics due to ignoring the covariates used in the minimization allocation. In this simulation we assume three binary (0 or 1) covariates, x_1 , x_2 , and x_3 , and they influence the survival time (end point) through a proportional hazard model:

$$\lambda(t; x) = \lambda_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3).$$

The distribution of survival time is assumed to be an exponential one (this is not an essential restriction), and right-censoring is not assumed. The number of strata is $2^3 = 8$, and for each strata 4, 8, and 16 patients are assumed. (Total number of patients is 32, 64, and 128.) The order of entry of patients is randomly permuted, and each patient is allocated deterministically into one of two arms with the minimum total of marginal frequency. The figures in the table are actual significance levels for null hypothesis. For β s, two sets of values (setting 1: 0.77, setting 2: 1.54) are assumed.

Conservativeness is clear from Table 1, and it is certain there is a power loss if the covariates are ignored. A little liberalism seen in the stratified analysis with a small number (4) of patients in each stratum may be a problem because of dependence between the numerator and the denominator of the Mantel-Haenszel type statistics, which is pointed out by Brown (9).

Permutation Tests and Development of a Computer Program

A design-based analysis for clinical trial data allocated by the dynamic balancing is theoretically possible even for the minimization method. In a deterministic case, simulation can be carried out by permuting the order of the entry of patients (10); in a probabilistic case where

the allocation is done using a biased coin, random numbers can be generated by fixing the order of the entry (11). Kalish and Begg (2), however, state that "these methods require specialized computer programming and we are unaware of their use in practice."

We have developed a computer program for permutation tests to test the difference of survival times in two independent groups. This program can cope with the following two designs: deterministic minimization with Zelen's option (12) that prohibits severe imbalance within each institution, and stratified permuted block design within each institution where the block size is less than or equal to 8. (Permutation is carried out by fixing the number of patients allocated to each group in each block.) At present, the limitation of the problem is as follows: number of institution is less than or equal to 50, number of the total of levels of covariates (stratifying variables) is less than or equal to 50, and number of patients is less than or equal to 500.

Statistics for which permutation distributions are calculated are the log-rank statistics and the Peto-Pren-tice-type generalized Wilcoxon statistics. For the minimization method, both stratified and unstratified statistics are calculated.

This program is written in FORTRAN 77, and the number of lines is about 1600 without comments. This program only outputs the simulation result into an external file; the analysis including calculation of p -values, tabulation, and graphic presentation is carried out by SAS. Examples of the execution time (by Hitachi M680 with about 20 MIPS) are included in Table 2. We are planning to carry out 5000 to 10000 iterations in practice, and the execution time is reasonable as well as realistic if we can use a high-speed computer. (Fortunately we can.)

Currently we have no real example of permutation tests for the minimization method because the trial using the minimization is now under way. Figure 1 is a permutation distribution of the log-rank statistics for real clinical trial data. In this trial, the treatment allocation was done using a permuted block design within each institute (block size: 4) and the total number of patients was 96. There were 45 blocks in all, only four of which were complete, and in 22, only one of the treatments was allocated. The Mantel-Haenszel variance of the unstratified log-rank statistics is 13.5 and the permutation variance is 9.5; the p -value of the former is 0.531 by normal approximation, and the p -value calculated from the permutation distribution is 0.486. The bias of the center of the distribution from 0 seen in Figure 1 is due to the imbalance of the treatment al-

Table 1. Conservativeness due to ignoring covariates in minimization. Type I error (actual significance level): number of iteration is 5000.*

N	Unstratified long-rank		Stratified log-rank	
	5%	10%	5%	10%
Setting 1				
32	3.5%	7.6	6.4 ^b	12.2 ^b
64	2.8	6.4	5.6	10.4
128	2.9	6.8	5.8	10.7
Setting 2				
32	1.0	2.8	5.9	11.3
64	0.7	2.2	5.8	10.6
128	0.5	2.3	5.5	10.6

*SE of simulation 1% = 0.20; 5% = 0.31; 10% = 0.42; N, total sample size.

^b May be a problem of Mantel-Haenszel procedure in small samples pointed out by Brown (9).

Table 2. Examples of the execution time.*

Examples	Patients	Iterations	Minutes
Permuted block	100	1,000	0.1
		100,000	12.0
Minimization	200	10,000	2.5
	60	1,000	4.0

*By Hitachi M680 with approximately 20 MIPS.

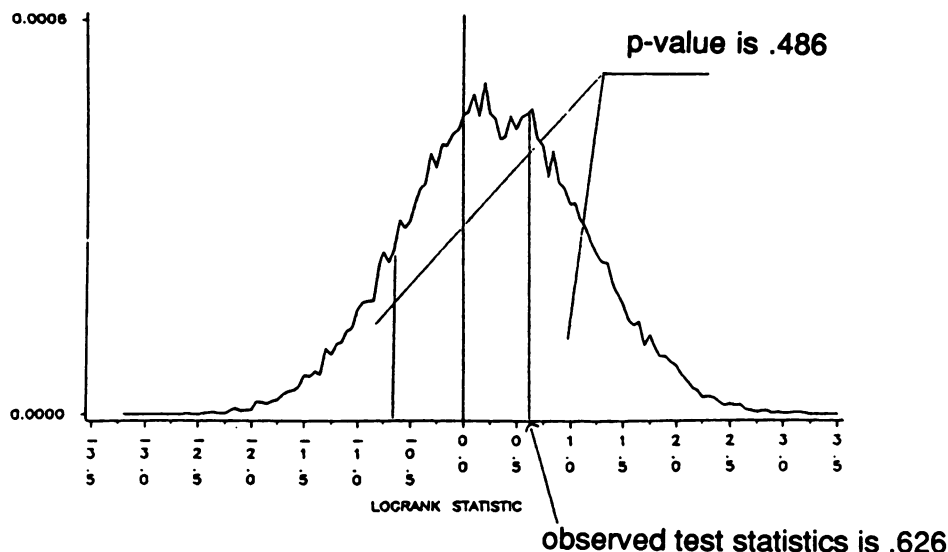


FIGURE 1. Permutation distribution of log-rank statistics (a real example: 96 patients, permuted block design).

location within blocks. In this example, the treatment effect is not significant in both analyses; but in a critical setting, a significant result may be derived by the permutation test, not by a simple analysis.

Discussion and a Suggestion

The validity of our permutation tests is based on the indifference of censoring patterns (distributions) in two treatment groups, and the check of this assumption is necessary in practice. For a deterministic minimization, there should not be a time trend in patients' response for the permutation test to be valid. When this assumption is doubtful, the probabilistic scheme using a biased coin should be adopted for avoiding possible biases.

We regard a permutation test as a kind of insurance; we expect that the result of a permutation test does not differ essentially from that of a stratified analysis, and the former reinforces the latter as well as the result of model-based analyses. But we should note that there is no theoretical justification for stratified analysis (especially, for its power), and relying only on the model-based analysis is sometimes hazardous in the credibility of the derived conclusion.

Data Management in Cancer Clinical Trials

Since 1985 an international group of researchers has been trying to establish an international minimal guideline (consensus) for clinical research in bladder cancer; the first international meeting was held in Antwerp in 1985 and the second, in Japan in 1987. The papers resulting from the first conference are compiled by Denis et al. (13).

One specific feature of this conference is that contributions from statisticians and data managers are expected and welcomed, and four working groups of 21 researchers are devoted to discussing the biostatistical/managerial problems in the second conference. The titles of the working groups are as follows: Statistical Analysis and Sample Size Determination, Determination of Prognostic Factors, Policy on Reporting and Publishing, and Data Management.

The fourth group consisted of three U.S./European data managers, three statisticians including the author, and several clinicians including three Japanese. The results of a long and earnest group discussion are summarized by De Pauw et al. (14). This working group presented the first opportunity between Japanese clinicians and U.S./European researchers to discuss ways of conducting cancer clinical trials and the organization for quality control. Japanese clinicians confessed that learning about a well-organized system of U.S./European clinical trials conducted in statistical (coordinating) centers was a kind of cultural shock because such information is difficult to get from research papers and absolutely different from their familiar Japanese system.

The working group found that there is diverse terminology for the same job (Table 3), there is no statistical center in Japan for cancer clinical trials, and organization in hospitals for collecting data and quality control ranges from a very elementary level to a highly sophisticated one from country to country. The group concluded that reaching a consensus on the organization of data collection and quality control is too ambitious and consensus should be on what should be done rather than by whom it should be done. The paper proposed a minimal guideline on the following items: protocol design; form design; collection of forms; computerization

Table 3. Job titles in cancer clinical trials.

Europe	U.S.	Japan
Study coordinator	Study coordinator Study chairman Protocol chairman	Chief investigator
Data manager (statistical center)	Data coordinator Data administrator Data manager	—
Local data manager (at the hospital)	Clinic coordinator Institutional data manager Research assistant Research study nurse	—
Statistician	Statistician	Controller

of information; feedback from the statistical center; and organization within the hospitals including the review of patient's history, cystoscopic procedure, protocol entry, protocol treatment, protocol follow-up, quality control, and form completion.

In the formal organization in Japanese clinical trials (Table 3), the titles of data manager and data coordinator are missing, and the same role is usually played by persons in pharmaceutical companies who visit clinicians' offices, collect data forms, and check the completeness of them. It is possible to classify cancer clinical trials from many viewpoints, and an important classification criterion is who supports the trial in data management and gives financial support. In Japan there are many scientifically valuable cancer clinical trials that are initiated by researchers themselves. When conducting such a trial, researchers are lucky if they can get support from a company because many troublesome activities for data management and quality control are carried out by the company. (Usually the company supports chemotherapy trials that use the drug they are selling.) The problem comes in a trial that is conducted without such support.

Three Japanese clinicians, Dr. K. Obata (Nagoya Second Red-Cross Hospital), Dr. T. Uyama (Shikoku Cancer Center Hospital), and Dr. Y. Matsumura (Okayama University Hospital), who attended the working group on data management and summarized the difference between the U.S./European system that was adopted in statistical centers and the Japanese system in cancer clinical trials that were initiated and conducted by researchers (4). Table 4 gives a summary and description of the present problems in data management and the quality of Japanese cancer clinical trials (statements in the parenthesis are comments by the author).

We think the direct import of U.S./European system is neither possible nor beneficial because there are great differences in the Japanese cultural background and those in U.S./European countries. Looking for the reliable and effective research system for cancer clinical trials, especially for data management and quality control, is a big assignment for Japanese biostatisticians involved in clinical trials, and such biostatistical input

Table 4. A summary of present problems in data management and the quality of Japanese cancer clinical trials, comparing Japanese methods with U.S./European methods.

U.S./European methods	Japanese methods
Design of data forms	
Forms are designed from a protocol by the study coordinator with the help of the experienced data manager.	Data forms are a part of the protocol and are designed independently. (Sometimes inconsistency exists between the forms and the documented protocol.)
Data completion	
Usually forms are completed by medical secretaries.	Forms are completed by doctors themselves. (Urgent prompts are necessary to have the doctors fill the forms, but no problems exists due to inexperienced secretaries filling out the forms.)
Collection of data forms	
Immediately after completion. There is an urgent prompt for forms from the data manager in the center.	Collection is done often after a committee of leading clinicians requests the data forms. (There may be a serious delay.)
Data check	
The data manager in the center is responsible for data checks and he/she asks the local data manager about questionable data so that the problems are settled at an early stage.	The doctor is asked about questionable data when they are summarized or analyzed.
Pathology	
Prepared pathological specimens are collected and the central pathologists gives the grading and staging.	All diagnoses are done by local pathologists, based on the published guidelines, but there is no external reviewing system.
Quality control in general	
Quality control is a collaborative work among clinicians, data managers, and statisticians. Procedures are documented explicitly.	The essential attitude in quality assurance is to trust the doctor. Quality control procedures are implicit, if any.

will contribute much to the quality of Japanese clinical research.

REFERENCES

1. Simon, R., and Wittes, R. E. Methodologic guidelines for reports of clinical trials. *Cancer Treat. Rep.* 69: 1-3 (1985).
2. Kalish, L. A., and Begg, C. B. Treatment allocation methods in clinical trials: a review. *Stat. in Med.* 4: 129-144 (1985).
3. Pocock, S. *Clinical Trials*, John Wiley and Sons, New York, 1983.
4. Ohashi, Y. Randomization in cancer clinical trials (in Japanese). *Proceedings of the Society for Survival Time Studies on Cancer Patients* 8: 99-111 (1988).
5. Sakamoto, J., Ichihashi, H., Koike, A., Nakazato, H., Ohashi, Y., Tamura, H., Fujii, T., and Ogawa, N. The minimization method in a median-size clinical trial [in Japanese]. *Jpn. Cancer Chemother.* 15: 1935-1942 (1988).
6. Green, S. B., and Byar, D. The effect of stratified randomization

- of size and power of statistical tests in clinical trials. *J. Chron. Dis.* 31: 445–454 (1978).
7. Forsythe, A. B., and Stitt, F. W. Randomization or minimization in the treatment assignment of patient trials. Technical Report No. 28, Health Science Computing Center, University of California at Los Angeles, Los Angeles, CA, 1977.
 8. Kalish, L. A., and Begg, C. B. The impact of treatment allocation procedures on nominal significance levels and bias. *Control. Clin. Trials* 8: 121–135 (1987).
 9. Brown, M. On the choice of variances for the log-rank test. *Biometrika* 71: 65–74 (1984).
 10. Simon, R. Restricted randomization designs in clinical trials. *Biometrics* 35: 503–512 (1979).
 11. Brown, B. W. Statistical controversies in the design of clinical trials—some personal views. *Control. Clin. Trials* 1: 13–27 (1980).
 12. Zelen, M. The randomization and stratification of patients to clinical trials. *J. Chron. Dis.* 27: 365–375 (1974).
 13. Denis, L., Nijima, T., Prout, G., and Schröder, F. H., Eds. *Developments in Bladder Cancer*. Alan R. Liss Inc., New York, 1986.
 14. De Pauw, M., Griffin, P., Ohashi, Y., Geboers, A., Hilgar, A., Obata, K., Uyama, T., Matsumura, Y., Rozek, S., Van Reijswoud, I., Birch, R., Barton, B., Hammond, B., and Staquet, M. Minimal guidelines for the organization of data collection and quality control in multicenter clinical trials. In: *Development in Bladder Cancer*, Vol. 2. Alan R. Liss Inc., New York, in press.